Recent Developments in the American Community Survey

Charles H. Alexander

U.S. Census Bureau
Washington, DC 20233

Presented to the Annual Meeting of the American Statistical Association (ASA), Dallas, TX, August 1998.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

KEY WORDS: Rolling sample, multiyear averages

I. BACKGROUND

The American Community Survey (ACS) is being developed by the U.S. Bureau of the Census to update, and eventually to replace, the decennial census "long form" survey. The ACS will cover the same topics as the long form, providing detailed economic, social and housing profiles of communities throughout the U.S. This paper gives updates about research on the ACS, with particular focus on our evolving understanding of how multiyear ACS data are likely to be used.

The ACS is a rolling sample survey (see Kish, 1998) contacting a different set of addresses each month until it "rolls" through the entire population. The addresses will be a systematic sample from a regularly updated "Master Address File" of all residential addresses, spread across all areas each month. The annual sample will be approximately 3 percent, cumulating to about 15 percent over a 5-year period. This compares to the 17 percent sample for the 1990 long form.

The ACS is a mail survey with telephone followup of all nonrespondents for whom a telephone number can be obtained, and personal visit followup of one-third of the remaining nonrespondents. Annual average estimates for each year will be available in July of the following year. Additional information on the ACS is available at www.census.gov.

The goals of the ACS are:

- to update census data used in federal funding allocations, at the state level based on estimates of level and change, and for smaller areas based mainly on estimates of level;
- to provide information for state and local decision-making, including updated values for the decennial census data that are currently used to describe and compare areas, and new information on trends and changes;
- to update census data used as an input to other Federal statistical programs, including uses in the sample design and weighting of household surveys, and in various statistical modeling projects.

The ACS is being introduced according to the following schedule:

- 1996-1998 A demonstration period in selected sites (4 sites in 1996, 8 in 1997, and 9 in 1998).
- 1999-2001 Thirty-seven "comparison sites" representing diverse areas around the country, with a 5 percent annual sample so that 1999-2001 averages can be compared to 2000 long form data for even small areas such as census tracts.
- 2000-2002 A national comparison sample with a 0.7 percent annual sample for further comparisons with the census long form. This sample will have some clustering in rural areas.
- 2003-later The full ACS with an unclustered systematic sample of about 3 million addresses per year, spread across all counties.
- 2010 There will be no long form survey attached to the decennial census, since the ACS data will have replaced it.

Although the ACS will have the same basic content and data collection modes as the census, there are some methodological differences. The most important is that the ACS data are collected throughout the year, including each person at his/her current residence at the time of interview. The census collects data in the few months after census day, counting each person at his/her "usual" residence as of census day.

There inevitably will be differences in the nature of "nonsampling error" in the two surveys, due to differences in coverage, interviewer training, unit nonresponse rates, and completeness of the collected data. These differences are discussed in Alexander (1997) and in some of the research described in the next section. Although these differences are important to methodologists in trying to understand and improve the quality of the survey, initial indications are that these differences will have relatively minor impact on users of the data.

II. GENERAL RESEARCH UPDATE

"Last year's" research on the ACS focused on initial reviews of data from the 1996 test sites by "local experts" familiar with those sites. This research focused on comparisons to the 1990 census and on examining basic survey performance measures such as response rates, sampling error measures, and coverage ratios. Although some improvements in the survey were suggested, there were no disturbing surprises. This kind of study will be repeated with 1997 data. The studies will go deeper in some respects, including comparison of the ACS estimates to administrative sources in some sites.

Some new research has been completed this year:

- There was a test comparing different reference periods for the income questions (Welniak and Posey, 1998). After reviewing the results, we will continue to ask about income for "the last 12 months" rather than "the last calendar year."
- There was a study of vacancy rates, which appear to be slightly lower for the ACS than for census procedures; several possible explanations were offered.
- Procedures for interviewing in group quarters were tested, and will be implemented for 1999.
- There was a small cognitive laboratory study of respondents' understanding of the ACS residence rules. This led to simplified instructions to define "current residence" and to ideas for future research.
- Based on analysis of the 1996 results, new questions on seasonal occupancy were added for 1998. These are needed to reconcile the ACS residence rule with intercensal population estimates.
- There was further work on weighting and variance estimation, using 1996 and 1997 ACS data, which was reported in Session 32 at these meetings.

III. MULTI-YEAR ACS DATA: BACKGROUND

The main focus of this paper is our evolving understanding of some of the issues related to using multiyear averages from the ACS. This includes 1) interpreting and using multiyear averages for describing and comparing areas, 2) use of multiyear averages to measure "need" for funding allocations, and 3) the likely uses of ACS data on changes and trends.

In the documentation for our data products, we have recommended that for basic descriptive statistics for small areas (or domains), data users should cumulate several years of data depending on the population of the area. Our recommendations are:

Table 1

Length of Average	Recommended Size Cutoff
1 year	$\geq$ 65,000 population
2 years	$\geq$ 30,000 population
3 years	$\geq$ 20,000 population
4 years	$\geq$ 15,000 population
5 years	< 15,000 population

The cutoffs correspond to a 12 percent coefficient of variation for a 10 percent estimate with a "typical" design effect. One-year estimates will be available for all size categories, but the averages are recommended to give sufficient precision.

This recommendation has been a source of concern for many potential users of ACS data, especially those most accustomed to point-in-time census data, who have asked how these averages are to be interpreted. One problem in addressing these concerns has been that we have never written down an explicit model for the statistical problem that we think users are addressing. That is a major focus of the remainder of the paper.

As we learn more about likely uses of ACS data, three major types of use can be distinguished:

- Basic description or comparison of areas
- Predicting current need, for possible use in allocating funds
- Measuring changes over time

As a preview of the remainder of the paper, for basic description the author still favors using averages with cutoffs such as those given in Table 1. For predicting current need, averages still make sense, but the initial results suggest more emphasis on 3-year averages even for small areas. For measuring changes over time, the user should in general analyze the time series of annual estimates, although averages are sometimes convenient as a simple compromise solution especially as an adjunct to a descriptive analysis that makes use of averages.

IV. MODELS FOR BASIC DESCRIPTIVE USES OF ACS DATA

There are three basic models under which an average of previous year's values is the "statistic of choice." Under these models, the data user would use the average, and interpret it as dictated by the model. If these models do not describe the user's view of the problem, then the user should start with the annual time series and analyze it as dictated by whatever other model is being adopted. The models are:

- the mean for a particular period of time is of interest
- the "typical census users" model
- a "random noise" model

Notation: To describe the three models, let us consider using either Census 2000 data or 1998-2002 data in the year 2003. Let X_t be the actual value of interest for some particular area in year t. The current year is t = 2003.

Rather than the actual value, what is observed is

$$\hat{X}_t = X_t + \varepsilon_t, \quad (4.1)$$

where ε_t is sampling error. For the census, $\hat{X}_t$ is observed only every tenth year, but has a smaller sampling variance than any single year of the ACS.

Model 1: The mean for a particular period of time is interest.

In this model, the assumed goal is to estimate

$$(X_{1998} + \ldots + X_{2002}) / 5$$

In this case, the ACS average is an obvious estimator, and can be justified under a variety of models for the time series $\{X_t\}$.

An example where this model would apply is when comparing the racial or ethnic distribution of bank loan recipients to the distribution for the surrounding community, which is needed as part of enforcing laws covering "fair lending practice." If the period of bank loans under consideration is 1998-2002, then it makes sense for the ACS descriptive data for the community to cover the same time period.

Although this is an important application, it is atypical. Typically the interest is in "the way things are now", as reflected in Model 2.

Model 2: "Typical census uses" model. As an example of a "typical" use of census data, consider the popular type of visual display in which a map shows census tracts, with shades of color indicating the percent of people or housing units in the tract with a given characteristic. For example the goal may be to display where in the county different ethnic groups "are" concentrated.

Our question is: if maps are used for that purpose in 2003 with 2000 census data, what assumptions are being made about changes over time, and how would the interpretation be different if the 1998-2002 average were used instead? This question equally well could be asked about the interpretation of tabular data, or other forms of display.

Based on questions raised by census users in sessions such as this one, the author hypotheses that the following model is being used. Perhaps this could be tested by appropriate "cognitive" surveys of census data users. The general idea behind this model is that 1) the interest is the way things "are"; 2) the census estimates are used as though they describe the current situation; 3) the user recognizes that the values for some areas may have increased or decreased since 2000; but 4) no explicit adjustments are made for such changes. More specifically, model 2 is:

- implicit assumption (default model)

$$X_{1998} = X_{1999} = \ldots = X_{2003} = \mu$$

- alternative 1 (trend)

$$X_t + X_{1998} + C\,(t\text{-}1998),$$

where $c \neq 0$

- alternative 2 (sudden jump)

$$X_t = X_{1998} \quad \text{for } t < J$$
$$X_t = X_{1998} + c \quad \text{for } t \geq J$$

The user views the map as though the default model is true, but has some concern about robustness under alternatives of the forms "trend" or "sudden jump." Different tracts may have different trends or jumps since the census, but it is not known which alternative applies to which area.

The following table gives the interpretation of the census estimate and the 5-year average under the various alternatives.

Table 2

	Census Estimate	5 - year average
Default Model	"The way things are"	"The way things are"
Trend	The way things were 3 years ago ("year 3")	The way things were 3 years ago ("year 3")
Sudden Jump in year 4, 5	The way things were before the jump	An average of before and after... more like before
Sudden Jump in year 1, 2	The way things are after the jump	An average of before and after... more like after

The interpretations are somewhat different in the case of a jump. For a single tract where the timing of the jump is known, the single-year census estimate is easier to interpret than the 5-year average. However, when there are many tracts and it is not known which have jumps or when the jumps occur, the situation seems similarly complex with either the census or the average. The ACS has the advantage that some information about which

tracts may have trends and jumps can be obtained by looking at the series of single-year estimates.

Model 3: Random noise model. Of course, changes may be more irregular than "trends" and "sudden jumps." In one extreme case of irregular changes, the average may still be the estimator of choice for different reasons than under Model 2. In particular, suppose that the actual value for the area of interest is

$$X_t = \mu + \eta_t$$

where $E(\eta_t) = 0$

and the η_t 's are uncorrelated.

This might describe a small town of 20 housing units where one year 3 households are in poverty, the next year 6, the next year 4, and so forth. In this case, the 5-year average is interpreted as an estimate of μ, and η is uninteresting "noise." (Note that η represents variations in the actual value in the population, as contrasted with ε which represented sampling error.)

Use of the 5-year average assumes that this model applies over a 5-year period. Over much longer periods, one would eventually expect trends or changes in μ, in addition to the noise.

V. USES OF AVERAGES TO ASSESS CURRENT NEED

Many of the formulas used to allocate funds from federal government programs make use of data from the most recent census long form. A variety of approaches are used. Some programs allocate funds to states, and state agencies separately distribute the funds within the state. Some formulas also make use of intercensal population estimates from demographic models, some use state-level estimates from national household surveys, and some use model-based estimates such as the Bureau of Labor Statistics' Local Area Unemployment Statistics or the Census Bureau's Small-area Income and Poverty Estimates for counties. A discussion of the wide variety of approaches used in these formulas is beyond the scope of this paper, except to note that the formulas are established either directly by Congress or by the federal agency administering the program, not by the Census Bureau.

The following is a simple model for the problem. The actual "need" for funds in the current year is measured by some variable X_t for a particular geographic area. For example X_t might be the number of children in poverty in that area. The available data are used to obtain an "assessment" of the current need:

$$\hat{A}_t = f(\ldots, \hat{X}_{t-i}, \ldots, \hat{X}_{t-1})$$

where f is some function and $\hat{X}_{t-i}$ is a sample estimate as in (4.1). This assumes that the current year estimate $\hat{X}_t$ is not available at the time the need is assessed. The time series distribution of $\{X_t\}$ is unknown for any given area and may have a very different models in different areas and for different variables.

If distribution of the time series $\{X_t\}$ could be determined for a particular area and variable, then an optimal forecasting function f might be determined. However, the optimal assessment f may vary by area. The average might be used as a simple "compromise" predictor for all areas.

The following alternatives will be considered in what follows:

1 year: $\hat{A}_t = \hat{X}_{t-1}$

3 year: $\hat{A}_t = 1/3\,(\hat{X}_{t-3} + \hat{X}_{t-2} + \hat{X}_{t-1})$

5 year: $\hat{A}_t = 1/5\,(\hat{X}_{t-5} + \ldots + \hat{X}_{t-1})$

Previous census:

$\hat{A}_t = \hat{X}_0$ is used for t=2, ..., 11,

where t=0 denotes the census year.

Many people have suggested that it would be better to use a weighted average with more weight on more recent years. The problem is how to achieve a consensus on the proper weights. It is possible that further research will suggest a clear answer, especially for specific purposes. In the meantime, we continue to recommend the "simple, familiar" unweighted average.

In comparing the assessed need $\hat{A}_t$ to the actual need X_t, there are two sources of error:

- Sampling error

- "Forecast bias", the difference between X_t and $E(\hat{A}_t)$

The partition into "sampling error" and forecast bias is very simple if a squared error less function is adopted. For example, if the loss function over an interval of years $[t_1, t_2]$ is

$$L(t_1, t_2) = (t_2 - t_1)^{-1} \sum_{t=t_1}^{t_2} (X_t - \hat{A}_t)^2$$

then the expected loss using a 3-year average is

$$E\left(L_{(3)}(t_1, t_2)\right) = (t_2 - t_1)^{-1} \sum_{\text{funct}=t_1}^{t_2} (X_t - 1/3(X_{t-3} + X_{t-2} + X_{t-1}))^2$$

$$+ (t_2 - t_1)^{-1} \sum_{t=t_1}^{t_2} (\text{Var}(\varepsilon_{t-3}) + \text{Var}(\varepsilon_{t-2}) + \text{Var}(\varepsilon_{t-1}))/9$$

The first component measures the difference between the current value of X_t and the average of the 3 previous values, assuming these values were measured without error. The second component is the sampling variance of the averages used to make the allocation. A similar partition applies to other lengths of averages and to estimates based on the previous census.

An important question is whether our criteria for using multiyear averages from Table 1, which were based only on considerations of sampling error, still apply to this forecasting problem.

A Simulation: A simple simulation illustrates how the two components interact. Table 3 shows the root mean square prediction error expressed as a percentage of the actual value, i.e.,

$$100 * E(X_t - \hat{A}_t) / X_t,$$

averaged over the 10-year period when a particular census's data would be used (years 2 through 11, where zero is the census year).

The simulation assumes the regular 3 percent ACS sampling rate for each year for the "ACS" data and the 17 percent 1990 overall census long form sampling rate for the "census" data. This table assumes a "typical" design effect of 2.0 for the census data and the "typical" ratio of 6.25 relating the variance of the ACS annual estimate to that of the corresponding 1990 long form estimate.

The population of the hypothetical area of interest is assumed to be 10,000 and X_0=1000. The rows of the table correspond to various assumptions about how X_t changes over time. A linear growth in X_t is assumed, varying from no change to an annual increase of 20% of the initial value (i.e., X_0=1000, X_1=1200, ..., X_{11}=3200).

Table 3

Area Population = 10,000 Population of interest = 1000 in year zero Table gives prediction error averaged over years 2-11, with linear growth				
Annual growth rate in group of interest	RMS Prediction Error (as percent of estimate)			
	Census	1-yr. ACS	3-yr. ACS	5-yr. ACS
0	9.2	25.7	14.9	11.5
1%	11.0	24.9	14.5	11.5
2%	14.5	24.2	14.3	12.0
3%	18.3	23.5	14.4	12.9
5%	25.2	22.4	14.9	15.1
7%	31.0	21.5	15.6	17.4
10%	38.2	20.5	16.9	20.9
20%	53.8	18.7	20.8	28.8

Note that if there is no growth, the census does best because of its lower sampling error, although the 5-year average comes close. Even a small amount of growth (2%) brings the 5-year average and even the 3-year average ahead of the census.

As far as the length of the average, for smaller areas like this one, the single year estimates are the worst unless there is a very high growth rate.

However, the 3-year and 5-year averages are fairly close overall: the 3-year average has moderately higher sampling error, but the 5-year average has a greater lag in picking up growth.

Discussion: There is no simple conclusion about whether the 3-year average or 5-year average is preferable, because different areas have different growth rates and we have not specified a loss function that allows the "losses" for different areas to be combined into an overall "loss." However, it is clear that considering forecast biases moves the optimum toward shorter averages. In this table the 3-year average certainly seems competitive with the 5-year average, and even for the comparable tables for areas of 5,000 population an argument could be made for the 3-year average, if larger errors are of greatest concern.

As the area size decreases much below 5,000 population, the pattern is more complicated since both the ACS and census long form have an oversample of small governmental units and small school districts, so the sample size does not decrease proportionally to the population size.

At this stage of the discussion it would be premature to make very specific conclusions about implications for funding formulas. However, it is clear that the appropriate criteria for the length of averages are not necessarily the same if the averages are used for funding formulas as when they are used for basic description of areas.

The limited results cited above raise the possibility of using 3-year averages for all sizes of areas. Those results by themselves would only justify using 3-year averages for smaller areas; indeed, considerations of forecast bias would even more strongly favor 1-year averages for larger areas. The arguments for using 3-year averages for larger areas would be i) the convenience of using the same length of average for all areas; ii) the 3-year averages would have less year-to-year change in the allocation, making funding more predictable and facilitating planning of how to use the funds. Some state-level funding formulas presently use 3-year averages for this reason.

An additional criterion: Besides the error in predicting individual year's need, as in (5.1), an important secondary criterion may be the difference between

$$\sum_{t=t_1}^{t_2} \hat{A}_t \quad \text{and} \quad \sum_{t=t_1}^{t_2} X_t$$

In other words, how does the total assessed need for a particular area over a period of years compare to the actual total need? This is particularly relevant if funds are to be allocated in exact proportion to the assessed need, but in any case it seems desirable for this difference to be small.

This criterion strongly favors updated averages over a decennial census. In the simplest case, where $\hat{A}_t = \hat{X}_{t-1}$, then looking at the total allocation over years "2 through 11", when a particular census would be used, the total allocation for the ACS is

$$\sum_{t=2}^{11} \hat{A}_t = \sum_{t=1}^{10} X_t + \sum_{t=1}^{10} \varepsilon_t$$

For the census, this would be

$$\sum_{t=2}^{11} \hat{A}_t = 10\, X_0 + \varepsilon_0$$

The sampling error terms both have expected value zero. Since typically a single year's ACS sampling error has variance $Var_{ACS}(\varepsilon_t) \approx 6.25\, Var_{LF}(\varepsilon_0)$, the sampling error component of (5.2) has variance on the order of 62.5% of that of (5.3).

The more important advantage of the ACS is seen when comparing the expected value of (5.2) and (5.3) to the target value of

$$\sum_{t=2}^{11} X_t .$$

For (5.2), nine of the terms in the summation

$$\sum_{t=1}^{10} X_t$$

are identical to those in the target summation. For (5.3), the expected value is based only on X_0.

If a time shift in the allocation is accepted, the agreement between the ACS single-year allocation and the target value is even more exact. The expectation of the total assessed value for years 3-12 is equal to

$$\sum_{t=2}^{11} X_t ,$$

which is equal to the desired allocation for years 2-11. In other words, in expected value the area gets exactly what it should get, just one year late.

For 3-year and 5-year averages, the agreement between

$$\sum_{t=1}^{11} \hat{A}_t \quad \text{and} \quad \sum_{t=1}^{11} X_t$$

tends to be less exact than for the 1-year value, although still better than for the census. Thus considering the total assessed need over a period of time further favors shorter averages.

VI. MEASURING CHANGES OVER TIME

As we have described the ACS to potential data users around the country, we have found a number of potential uses of the capability of measuring changes over time. Table 4 illustrates the precision of the ACS for measuring changes from one year to the next.

There are important uses of year-to-year change estimates at the state level relating to such public issues as welfare reform. We have also found interest from local planning agencies in knowing about steady trends in smaller areas, which can be described by moving averages. An example is growth in the number of children age 0-5 "not speaking English well," for educational planning.

Very large changes from one year to the next in small areas can also be measured, but not very precisely. For example, in rural areas with large new housing developments, it would be possible to see where the growth occurred and get an idea of the characteristics of the new housing and its residents. However caution in interpreting the results would be required because of the high standard errors. The popularity of such analyses of "noisy" year-to-year changes for small areas remains to be determined.

Table 4

Measured Increase from One Year to the Next Required to be Statistically Significant (Two-Sided Test with α = 1)		
Population	Rate in Year 1 (percent)	Min. Significant Year 2 Value
500,000	10.0	11.0
250,000	10.0	11.4
100,000	10.0	12.2
65,000	10.0	12.7
30,000	10.0	14.0
20,000	10.0	14.9
15,000	10.0	15.7
5,000	10.0	20.0

In general, the time series of annual data should be the starting place for measuring changes over time. In some cases, moving averages would be used to "track trends," i.e., to smooth the time series so that trends can be seen at the expense of higher frequency movements in the time series. The cutoffs in Table 1 may be a convenient place to start for this purpose, but the optimum length of the moving average depends on both the size of the area and the length of the "trends" that are of interest.

VII. VARIABLES WHOSE "MEANING" CHANGES OVER TIME

Another concern about multiyear averages is how they work for variables whose "meaning" changes over time. Some examples:

- number of people who lived in a different county 5 years ago
- income (because of inflation)
- age (because year of birth changes)

The reference date for "5 years ago" is different for different years in the 5-year period. For the 1998-2002 averages, the characteristics of the set of people who lived in another county in 1993 may be different than the characteristics of the set who lived elsewhere in 1997. The "meaning" of a dollar changes over the 5 years because of inflation. The 20-year-olds in 1993 have a different birth cohort than those in 1997.

There are three basic alternatives for dealing with examples like these:

Focus on invariant aspects of the category. In other words, ignore the issue. For example, define a "recent in-migrant" to be someone who lived in a different county 5 years ago. Then the first example is just looking at the variable "number of recent in-migrants", whether for 1998 or 2003. The distribution of characteristics of recent in-migrants may change over the years, but that may be true for any category of any variable, even for fixed characteristics such as gender. This gets back to the issues discussed in Section IV, where now "category" or "domain" takes the place of "area."

Inflation adjustments. For dollar-denominated variables, we think inflation adjustments would be made before averaging years. Note that multiyear frequency distributions are computed by averaging the various year's estimated numbers in each cell of the distribution. For the 1996 ACS annual estimates, inflation adjustments were applied to months during the year before producing the annual income distribution, although this was not vital during this period of low inflation. Research on income data from the ACS continues. (Welniak and Posey (1998).)

Use annual data and include time as a variable. In the third example, if birth year is an important variable, then the model should start with annual data and include birth year in the model. Analysis of the model may indicate that the time variable can be "collapsed," and if so the final analysis may use multiyear averages. However, if time is an important variable, then the data for individual years would be needed.

VIII. OTHER RESEARCH PLANNED OR IN PROGRESS

In progress:

- Closer integration of ACS and intercensal demographic estimates
- Use of aggregate administrative records data to improve ACS poverty estimates
- Use of ACS data in other statistical modeling programs
- Within-household coverage
- Rostering and other questionnaire issues
- Improved access to the ACS data (e.g., new CD-ROM)
- Cost modeling and operational issues
- "Smoothed" formula for sampling rates as a function of population for oversampling small governmental units
- Tests of intercensal address-list updating method

Planned:

- Close study of differences between 1999-2001 ACS and 2000 long form in comparison areas
- Impact of ACS in 1998 census dress rehearsal site
- "Adaptive sampling rates" in low-response areas
- Developing a policy for adding supplemental questions to the ACS

An upcoming workshop sponsored by the Committee on National Statistics of the National Science Foundation will review the ACS research plans and may suggest additional topics.

References

Alexander, C.H. (1997). " The American Community Survey: Design Issues and Initial Test Results." To appear in Proceedings of the XIV Annual International Symposium on Methodology Issues. Statistics Canada.

Kish, L. (1998). " Space/Time Variations and Rolling Samples." *Journal of Official Statistics*, Vol. 14, No. 1, pp. 31-46.

Welniak, E. and Posey, L. (1998). "Income in the ACS: Comparisons to the 1980 Census." Presented at the ACS Research Symposium, March 25, 1998, Bureau of the Census. To appear on the ACS website at www.census.gov.